# Part 3 - Session Papers for the EPA 22nd Annual National Conference on Managing Environmental Quality Systems

**April 14-17, 2003 New Orleans, Louisiana**

---

## ENVIRONMENTAL STATISTICS AND PUBLIC HEALTH

**Investigating Cancer Mortality and Environmental Information** - M. Conomos, U.S. EPA

**Using Small Area Analysis to Estimate Cumulative Prevalence, One Year Period Prevalence, and One Year Severe Period Prevalence of Asthma in Chicago Public Schools** - T. Brody, U.S. EPA

**Depiction of Population Characteristics in Areas Impacted by Industrial Source Emissions** - L. Lehrman, U.S. EPA; A. Lubin, U.S. EPA

## EMISSIONS, EXPOSURES, AND RISK ASSESSMENT

**Establishing Hazardous Emissions Limitations: Taming a Modern Hydra using Statistical Methodology** - N. Andrews, U.S. EPA

**Estimation of the Exposure Point Concentration Term Using a Gamma Distribution** - A. Singh, Lockheed Martin

**Integration of Spatial Data: Methods Evaluation with Regard to Data Issues and Assessment Questions *** - Michael O'Connell, U.S. EPA

* **Abstract Only**

# Investigating Cancer Mortality And Environmental Information

Margaret G. Conomos, Ruth H. Allen, Rashmi Lal, Maryann Suero, Cheng Ling
U.S. Environmental Protection Agency, Washington, D.C. 20460-0001

*The objective is this investigation is to better understand the relationship between cancer and environmental pollutants. From the National Cancer Institute (NCI), we chose the cancer mortality data that is available on their Internet site as the Cancer Atlas. For environmental pollutant information, we chose to focus on a database that is also widely used and accessible: the Toxic Release Inventory (TRI) available on the internet as TRI Explorer. This ecological study initially explored derived environmental statistics from counties in Northeast Ohio with high rates of childhood leukemia over more than 40 years and from toxic release data spanning more than a decade. Exploratory data analysis suggests that higher releases of chemicals in the past are associated with observed higher cancer mortality. Additional regions explored include all counties of all states with high cancer mortality rates and/or large air emissions of cancer causing chemicals. Further analysis comparing and ranking all states does not add to the association. The strengths and limitations of an ecological study will be discussed. The integration of health and environmental data is in its infancy. The human environment, unlike a laboratory experiment, has multiple cancer causing agents and a wide range of differing human susceptibilities due to inherent genetic variation and lifestyle choices, besides the component of environmental exposures.*

**a. Introduction:** The objective of this ecologic study is to better understand the relationship between childhood leukemia and environmental pollutants. The mission of the U.S. Environmental Protection Agency (EPA) is to protect human health and to safeguard the natural environment — air, water, and land — upon which life depends. EPA receives numerous reports on environmental information, and under several legal mandates regulates the releases of pesticides and toxic materials into the environment, such as the air, water, and land. Over time these releases have been reduced. The question is can we demonstrate that we have protected the public health in the process? The better we can understand relationships between diseases and environmental conditions, the better we can direct our efforts to prevent adverse impacts to the environment and people. Many investigators have approached this problem with the use of epidemiological methodologies, such as the case-control study or the cohort study design. We chose a practical exploratory analysis and ecologic approach that would allow us to compare reported information that is available in public health and environmental databases. EPA collects monitoring and reported release data on air quality, water quality, pesticides and toxic substances. We chose to focus on a database that is widely used and accessible: the Toxic Release Inventory (TRI) available on the Internet as TRI Explorer http://www.epa.gov/triexplorer. From the National Cancer Institute, we chose the cancer mortality data that is available on their Internet site as the Cancer Atlas. To explore our hypothesis, we needed to select a locality as a pilot area and we chose to study a region in Northeast Ohio that had adequate health and environmental data.

***b. Subjects and Methods: Subjects:*** We chose to study children, ages 0-19 years, for the primary reason that for a chemical to cause cancer that appears in childhood it must operate with

a relatively short latency period, or time from exposure to onset of disease. Also, children do not have the same opportunity to develop the lifestyle choices that increase risk of cancer in adults.[1] Moreover, we were challenged by Dr. Lynn Goldman, formerly of EPA who said, "We know little about how chemicals in the environment relate to risks of childhood cancer." [2] The literature suggests that the exposure of parents also can influence the occurrence of cancer in children.[3] The EPA has an active Children's Health Program, and ten regional offices. Each EPA Regional Office is responsible for the execution of the Agency's programs in particular States. We chose to study two counties in Northeast Ohio, which is among the six EPA Region 5 states. Region 5 serves Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin and 35 Tribes. We collaborated with Region 5, and their Children's Health Program, due to a shared interest in development of environmental indicators and pesticides in schools. Several considerations included high cancer mortality rates that have been associated with Northeast Ohio. Also, there are sufficient numbers of childhood leukemia cases available to analyze, as well as, programmatic interest, and high rates of industrial releases for known and suspected toxic chemicals, including solvents known to cause leukemia based on rodent tests.

***Criteria for Chemical Selection:*** We chose to investigate leukemia, and we identified leukemia causing compounds that were tested and reported in The National Toxicology Program (NTP) Report on Carcinogens (RoC) *http://ehp.niehs.nih.gov/roc/toc9.html*. The law stipulates that the Secretary of the Department of Health and Human Services (DHHS) shall publish a biennial report that contains a list of all substances (1) which either are known to be human carcinogens, the A list, or may reasonably be anticipated to be human carcinogens, the B list, and (2) to which a significant number of persons residing in the United States are exposed. We used the Ninth Report on Carcinogens (2001), also referred to as the NTP list, to identify substances that may cause leukemia in humans. We selected TRI chemicals that have been reported, starting in 1988 with the first full year of reporting, as being released at greater than 10,000 pounds in the geographic area of interest. We recognize that not all chemicals of interest have been tested and that release does not equal measured exposure or internal dose to an individual. Using the Internet, from the entire NTP RoC list of 213 substances, we identified 40 carcinogens, 12 known and 28 anticipated, which cause leukemia. We then went to the TRI Explorer that is a Web-based analytical tool that enables data users of the Toxics Release Inventory (TRI) to compile their own reports online. Since 1987, the TRI program has provided information to the public on releases and other waste management information for more than 600 chemicals and chemical categories from many industry sectors. We looked at TRI data for toxic chemicals released in Ashtabula and Summit Counties in Ohio. We narrowed our focus to carcinogens that cause leukemia from the NTP list, and to those chemicals that were also reported to be released in the counties of interest.

***Exploratory Spatial Analysis:*** The counties of Ashtabula and Summit, which include the city of Akron, in Northwest Ohio are well known to EPA. The Cuyahoga River flows for 100 miles,

---

[1] Steingraber S. Living Downstream: An Ecologist Looks at Cancer and the Environment. New York: Addison-Wesley Pub. 1997. p. 39.

[2] Goldman LR. Chemicals and children's environment: What we don't know about risks. Environ Health Perspec 1998; 106(Suppl 3): 875-880. p. 875

[3] Colt JS and Blair A. Parental occupational exposures and risk of childhood cancer. Environ Health Perspectives 1998; 106(suppl 3): 909-25.

through the cities of Akron in Summit County and Cleveland in Cuyahoga County before emptying into Lake Erie. The Cuyahoga River supports one of the most densely populated and industrialized urban areas in America. In 1969, the Cuyahoga became a stark symbol of water pollution when oil slicks on the river's surface caught fire. The burning river captured the attention of the nation and became a rallying point for passage of the Clean Water Act. In 1972, the cover of Time magazine announced that Lake Erie was dead. Also, in 1970, the White House and Congress worked together to establish the EPA in response to the growing public demand for cleaner water, air, and land. Through new laws and new partnerships, water quality on the river has improved dramatically, and slowly environmental releases were reduced. Rivers and the communities they support are now experiencing a rebirth.

Demographically, Ashtabula and Summit Counties in Ohio have comparable age distributions for persons under 18 years old, 26.2% and 25.0% respectively, and for females, 51.3% and 51.8%, respectively, based on the 2000 Census. Ashtabula County is more rural, with a land area of 702 square miles and a population of 102,728 and a population density of 146.2 persons per square mile. Black or African American persons represent 3.2% of the population. Summit County is much more urban, with a land area almost half that of Ashtabula, 413 square miles, and five times the population, with 544,217 people and a population density ten times greater of 1,315.4 persons per square mile. Black or African American persons represent 13.2% of the population.

***Cancer Maps:*** We used the Cancer Mortality Maps & Graphs Web Site, previously mentioned, that was developed and is maintained by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) *http://www3.cancer.gov/atlasplus*. The Web Site provides interactive maps, graphs (which are accessible to the blind and visually-impaired), text, tables and figures showing geographic patterns and time trends of cancer death rates for the time period 1950-1994 for more than 40 cancers. The site is based on data obtained from the National Center for Health Statistics (NCHS), the Federal Government's principal vital and health statistics agency. This site has several interactive data visualization tools to enhance the ability to view the data. This site encourages the study of geographic patterns of cancer that may provide important clues to the causes of cancer and improvements in cancer control.

Users of the Atlas will note that the cancer mortality data are not always available for all counties. Therefore, State Economic Areas (SEAs) combine counties, as a function of small numbers for some cancers. The maps from the Cancer Atlas show the geographical pattern throughout Ohio for leukemia. High rates for leukemia are in Summit County and Ashtabula for white males and females, and black males, ranging from 2 to 3.67 deaths per 100,000, age adjusted were observed.
There are sparse data for black females.

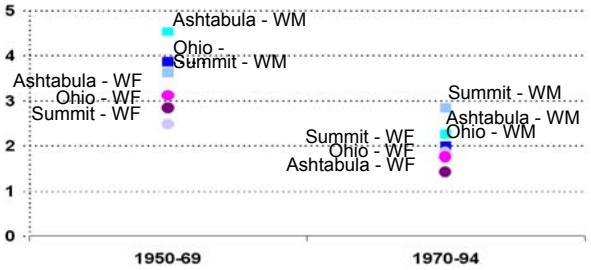



**Figure 1. Ohio childhood leukemia mortality rates/ 100,000 (county & state) for two time periods, ages (0-19).**

For young people in the age group less than 19 in Ohio, leukemia ranks 1st and has a rate between 1 and 2 per 100,000, and in the time of 1970-1994 resulted in 765 deaths in white males, 529 deaths in white females, 73 deaths in black males, and 51 deaths in black females. The childhood leukemia mortality rates in Summit and Ashtabula, as shown in Figure 1 have gone down over time. The rates were higher for white males and females in Ashtabula than Summit during 1950-1969, and reversed in 1970-1994. Most county rates were higher than the state rates.

The strengths of the Atlas include that it is a unique resource to help investigators identify patterns of mortality at the county and SEA level. Also, for cancers with poor survival rates and clear-cut diagnoses, mortality data closely reflect incidence data. For leukemia, this was true in the past, but the survival rates for children are dramatically improving.[4] Also, information from death certificates regarding the underlying cause of death is likely to be fairly accurate, because cancers are generally more accurately reported than other causes of death.

One limitation of the Atlas cancer data, however, is that often it is difficult to evaluate whether high mortality rates for less fatal cancers point to risk factors or indicate poor quality of health care. In addition, certain cancers, such as liver, brain, lung and bone, which are often the site of secondary metastases, may be incorrectly specified as the primary cause of death. Another limitation of the mortality data is that it is not possible to evaluate the effect that moving from one part of the country to another has on the death rates because residential histories are not included on death certificates. Also, for a younger age group, smaller racial group, or a smaller county for a rare cancer, the year to year percent variation will be greater than for all age groups combined or all counties combined, due to the smaller number of cases.

Note that the SEA cancer mortality data for Summit can be compared to the TRI county release data, because the Summit SEA and County boundaries are the same. However, the rural Ashtabula SEA is larger than the county boundary, but the grouped counties have similar economic and cultural characteristics.

***Toxicology and Environmental Release Data on Regulated Chemicals:*** The National Toxicology Program or NTP lists 213 chemicals that are either known, the A list, or anticipated, the B list, to cause cancer. The NTP lists 40 chemicals that are either known or anticipated to cause leukemia. Separately, we extracted from TRI Explorer all the chemicals released onsite.

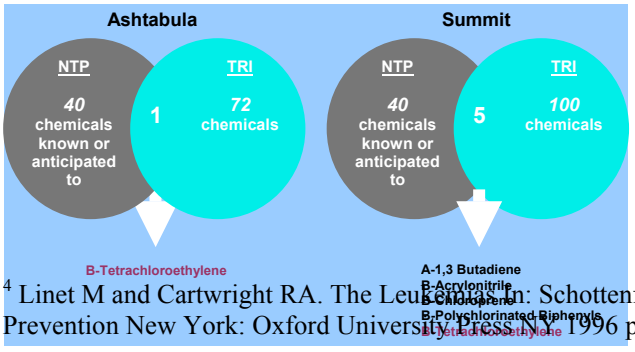



**Figure 2.** For 1988-2000, for Ashtabula County, the TRI list totaled 72 chemicals; for Summit County, the TRI list totaled 100 chemicals. We compared these TRI releases to the NTP list. For Ashtabula County, we found one TRI chemical released that may

[4] Linet M and Cartwright RA. The Leukemias. In: Schottenfeld D and Fraumeni JF, Jr. Cancer Epidemiology and Prevention New York: Oxford University Press, NY 1996 pp. 841-892.

cause leukemia: which is tetrachloroethylene. For Summit County, we found 5 TRI chemicals released that may cause leukemia: 1,3 butadiene, acrylonitrile, chloroprene, polychlorinated biphenyls, and tetrachloroethylene.

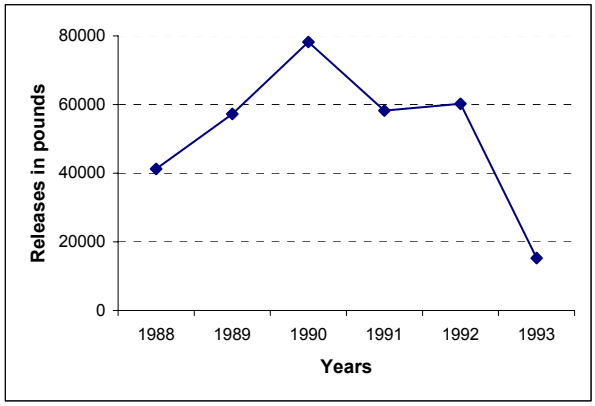



**Figure 3. Tetrachloroethylene on-site releases in Ashtabula County, Ohio**

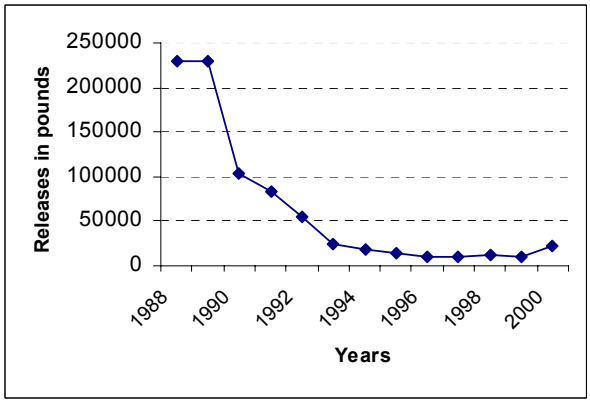



**Figure 4. 1,3 Butadiene on-site releases from TRI Explorer for Summit County,**

Figure 3 shows for Ashtabula County, tetrachloroethylene on-site releases were reported from 1988 to 1993, peaked in 1990 at about 80,000 pounds, and declined to 15,000 pounds in 1993. No on-site releases for tetrachloroethylene were reported after 1993. In Summit County, of the five chemicals causing leukemia that were released, two were not plotted. Polychlorinated biphenyls and chloroprene were released from facilities, but they were not released on-site. Figure 4 shows 1,3 butadiene on-site releases were reported from 1988 to 2000, peaked in 1988 and 1989 at about 230,000 pounds, and declined to approximately 20,000 pounds by 1994 and stayed at that low level from 1994 to 2000.

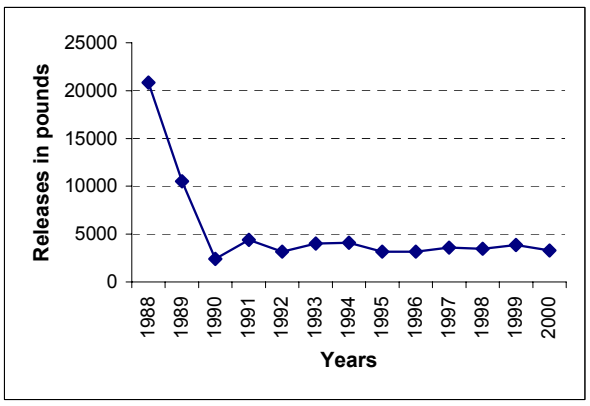



**Figure 5. Acrylonitrile on-site releases in TRI Explorer for Summit County, Ohio.**

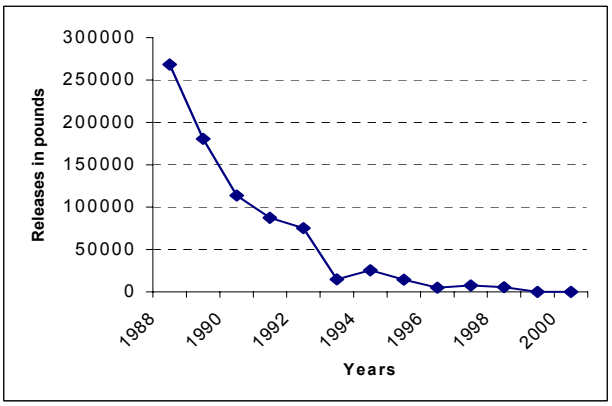



**Figure 6. Tetrachloroethylene on-site releases from TRI Explorer for Summit County, Ohio.**

Figure 5 shows for Summit County, acrylonitrile on-site releases were reported from 1988 to 2000, peaked in 1988 at about 21,000 pounds, and declined to less than 3,000 pounds in 1990 and stayed approximately at that level from 1991 to 2000. Figure 6 shows for Summit County, tetrachloroethylene on-site releases were reported from 1988 to 2000, peaked in 1988 at about

268,000 pounds, and declined to less than 20,000 pounds in 1993 and 1994, and declined further to 10 pounds in 2000. If we could plot release data prior to 1988, we expect the release data would be higher, based on the nature of the curve.
The TRI Explorer is a unique resource that enables TRI data users to compile their own reports online. However, it does not cover all industries. It covers: manufacturing sector (SIC 20-39), metal mining, coal mining, electrical utilities, chemicals and allied products wholesale distributors, petroleum bulk terminals and solvent recovery services. All facilities in the covered sectors do not report to the TRI Program and there are size requirements for reporting: the facilities which have 10 or more full time equivalent employees; and manufacture or process more than 25,000 pounds or otherwise use more than 10,000 pounds of any listed chemicals during the calendar year are required to report. Facilities often report estimated data and are not mandated to monitor their releases by TRI Program. The data are self-reported. Neither area sources nor mobile sources are reported in TRI and these have been shown to contribute significantly to air pollution.

***c. Discussion:*** Subjects are drawn from publicly available cancer mortality data because Ohio is not a part of the Surveillance, Epidemiology and End Results (SEER) cancer incidence data. Prevention programs would benefit by having access to cancer incidence data, and tracking pattern differences to measures the effectiveness of public health interventions, either in health care delivery, or in pollution abatement, especially in areas with high numbers of susceptible individuals. In the future, as state cancer incidence data become available from the Centers for Disease Control and Prevention (CDC) in a SEER-compatible form, a geographical information analysis approach can help to integrate available health and environmental data.

We found on the National Toxicology Program list many agents besides chemicals that cause leukemia, such as soot, tars, and pharmacologic agents. After reviewing the 40 substances listed which are known or anticipated to cause leukemia, the TRI program also listed very few of these agents. For the chemicals that were in common for both the NTP and the TRI program, larger amounts of the chemicals were released in the early years of reporting, 1988 to 1989. The amounts of the releases decreased over time. The higher releases of chemicals in the past may explain the higher cancer mortality that we observed. This case study began without prior knowledge of an association between leukemia and environmental pollution, and other study designs are needed to confirm such associations.

Reductions in the rates of childhood leukemia over the last fifty years are one of the notable public health intervention success stories. As with other studies which investigated environmental parameters and leukemia, a strong association was not evident, nor was it absent.[5, 6] Also, therapeutic improvements during the last few decades have been strongly associated with lower cancer mortality rates for childhood leukemia. [7] We used an ecologic study design with available records at the websites of the EPA and NCI. This approach has several limitations,

[5] Infante-Rivard C, Olson E, Jacques L, Ayotte P. Drinking water contaminants and childhood leukemia. Epidemiology 2000; 12: 13-19.
[6] Reynolds P, Von Behren J, Gunier R, Goldberg D, Hertz A, Harnly M. Childhood Cancer and Agricultural Persticide Use: An Ecologic Study in California. Environmental Health Perspectives 2002; 110: 319-324.
[7] La Vecchia C, Levi F, Lucchini F, Lagiou P, Trichopoulos D, Negri E. Trends in childhood cancer mortality as indicators of the quality of medical care in the developed world. Cancer 1998; 83: 2223-2227.

including no information on potential confounding variables, residential migration, and exposure assessment. However, we did not have to deal with recall bias. For future analyses, it would be more robust to pool the cancer mortality data for the two counties and compute the disease rate for the one two-county region and compare it to the rest of the state across time.

# Using Small Area Analysis to Estimate Cumulative Prevalence, One Year Period Prevalence, and One Year Severe Period Prevalence of Asthma in Chicago Public Schools

Thomas M. Brody, Ph.D.

*Small area analysis statistics can be used to estimate asthma prevalence in schools from national health data. In a pilot study, 1998 National Health Interview Survey data were stratified by age, race, and gender and projected onto Chicago Public Schools demography from the Department of Education's Common Core of Data for the same year. The estimates were brought into an address matched Geographic Information System of schools for visual interpretation. An initiative is under way to estimate asthma prevalence in a national set of schools while promoting the need for improved asthma surveillance systems.*

## Background

In 1998, Dr. Sandra Thomas of the City of Chicago Department of Public Health used childhood asthma statistics from the Center for Disease Control's 1993 National Health Interview Survey (NHIS) and 1990 census data for Chicago neighborhoods to estimate the asthma burden in the neighborhoods. Dr. Thomas presented the results in tabular form, which generated a great deal of interest, but it was clear that mapping the results in a Geographic Information System (GIS) would add much more meaning to the data. With Dr. Thomas's permission, the data were mapped in a GIS with the results shown in Figure 1. For the first time, residents of Chicago could view estimates of the spatial distribution of the asthma epidemic in their neighborhoods (Geist, 1999).

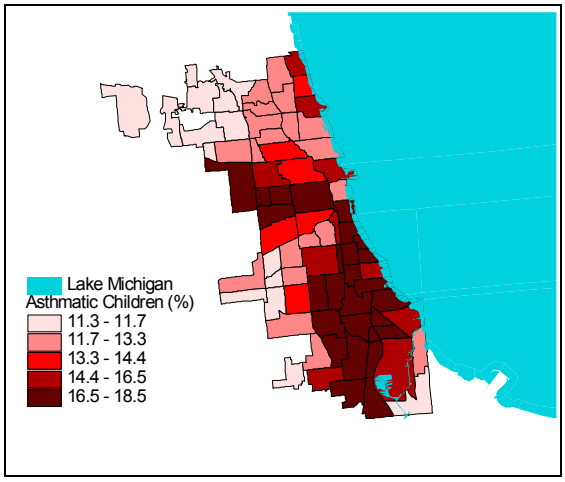



**Figure 1:** Estimated Asthma Prevalence Rate In 1993 In Chicago Neighborhoods For Populations Under the Age of 18.

The product was not perfect though. Important demographics had changed in the three years between the Census and NHIS data collection, and nine years between the Census collection and the publication of Dr. Thomas's estimates. Furthermore, the neighborhood scale looked somewhat awkward with color present in areas that lacked populations such as the Lake Calumet area to the south and the industrial corridor along Interstate 55 to the west. On some maps, like the one in Figure 1, populations were shown to be at O'Hare Airport on the northwestern tip of the city because the area is technically part of a neighborhood. Clearly, the specificity of the GIS needed to be improved and the results updated.

In 2000, Region 5 created a Children's Health Team to support the national Office of Children's Health Protection. Since many Children's Health programs were school based, the location of

schools was one of the team's first requests. The search for a school database culminated in the discovery of the US Department of Education's Common Core Data (CCD) file. This data set had a complete annual public school and biennial private school census of children broken down by gender, grade, and race. Dr. Thomas used similar demography for her analysis, so a new attempt was made to synthesize the 1998 NHIS statistics with the 1998 CCD.

## Method

Small Area Analysis is the term often used in the literature to define the broad range of methods that use data sources, statistical techniques, and computer applications to determine the number of events occurring in an area through comparisons with similar or larger area benchmarks (WIDHFS, 2003). Typically, small areas refer to county and sub county areas like cities, zip code areas, census tracts, or even smaller units (Murdock and Ellis, 1991). A recent text by Siegel provides an excellent review of these methods and their uses in describing health characteristics (Siegel, 2002).

The Small Area Analysis model used in this work is referred to in the literature as the synthetic method. The synthetic method creates an estimate of the population having a health characteristic in a small area by applying proportions of the population having the health characteristic in one or more demographic categories (age, sex, race, etc.) in a larger area to population figures for these demographic categories in the small area (ibid, 497). In this analysis, demographic proportions from the Sample Child Person (SCP) Section of the NHIS were synthesized with the same demographic categories in the CCD to estimate the burden of asthma in Chicago schools.

In 1998, the SCP contained a nationwide sample of 13,643 children. In this sample, 1,629 of the children had been told that they have asthma in the past (cumulative prevalence), 740 had had an episode of asthma in the past year (one year period prevalence), and 270 needed to visit the emergency room during a previous year's attack (one year severe period prevalence). Each of these prevalence types was stratified by the demographic categories of grade, race, and gender. Specifically, proportions of the population having cumulative, one year period, and severe one year period prevalence of asthma were calculated for: Male and Female; White, Black, Hispanic, Native American, and Asian; and grades Pre-Kindergarten through 12. The question of grade in the SCP survey asked for the highest level of school completed, so each given grade represented a person in the grade above. Unfortunately, that meant that pre-kindergarteners and kindergarteners had to be aggregated to create a single rate.

Additionally, it was thought that the model should include a sense of geography in order to account for the environmental features of Chicago. The most geographically specific information in the SCP is the subset of persons that live in a Consolidated Metropolitan Statistical Areas (CMSAs) above 250,000 in population in the Midwest Region. The CMSAs in this set include Chicago, as well as Indianapolis, Wichita, Detroit, Minneapolis/St. Paul, St. Louis (MO), Kansas City (MO), Omaha, Toledo, Cleveland, Columbus, Cincinnati, and Milwaukee. These areas all have similar climate, terrain, and population density, so it was assumed that proportions taken to reflect the geography of these cities reflected the same proportions for Chicago alone.

In the model, the universe of rates of each stratified demographic category and the population in each school in the stratified demographic category were multiplied together to create the asthma population in the school in the stratified demographic category. The resulting three total universes (grade, race, and gender) were divided by three assuming that each category independently affects an asthma outcome. Finally, the resulting population was multiplied by the ratio of the asthma rate from the Consolidated Metropolitan Statistical Areas (CMSAs) above 250,000 in population in the Midwest Region over the national asthma rates in order to weigh the outcome with more specific location based information.

Mathematically, the model can be written by letting

$g_i$ = the national asthma rate by grade $g$ for each grade $i$,
$s_j$ = the national asthma rate by sex $s$ in for each sex $j$,
$r_k$ = the national asthma rate by race $r$ in for each race $k$,
$c$ = the asthma rate for a large Midwestern CMSA,
$n$ = the national asthma rate,
$x_i$ = the number of students $w$ for each grade $i$,
$y_j$ = the number of students $x$ for each sex $j$,
$z_k$ = the number of students $y$ for each race $k$.

Then for each school, each type of prevalence $T$ is calculated by letting

$$T = \frac{c}{n}\left( \frac{\sum_i g_i x_i + \sum_j s_j y_j + \sum_k r_k z_k}{3} \right).$$

After the estimates were derived, the results were placed in a GIS containing the address matched Chicago schools from the CCD. The Chicago records in the CCD were matched to addresses using ArcView 3.2 and the Wessex 6.0 data set. The program address matched approximately 90% of the data, with 10% of the street segments missing from Wessex. The additional addresses were individually matched using Mapblast at www.mapblast.com. Finally, the asthma estimates were joined with the 593 public schools in the GIS.

## Results

The stratified rates are shown in Figure 2. Interestingly, the fact that the population was in a CMSA above 250,000 in population in the Midwest Region decreased the cumulative and one year period prevalence rates while only slightly increasing the severe period prevalence rates when compared to the national rates as shown by comparing "MSA" with "Nation." Other stratifications indicated that males have higher asthma rates than females; Native Americans have the highest cumulative and one year period prevalence rates of asthma, followed by Blacks, Whites, Hispanics, and Asians. However, Blacks surpass Native Americans and Hispanics surpass Whites in severe period prevalence of asthma. The cumulative and one year period prevalence rates increase through the school years, peaking in the seventh grade and then

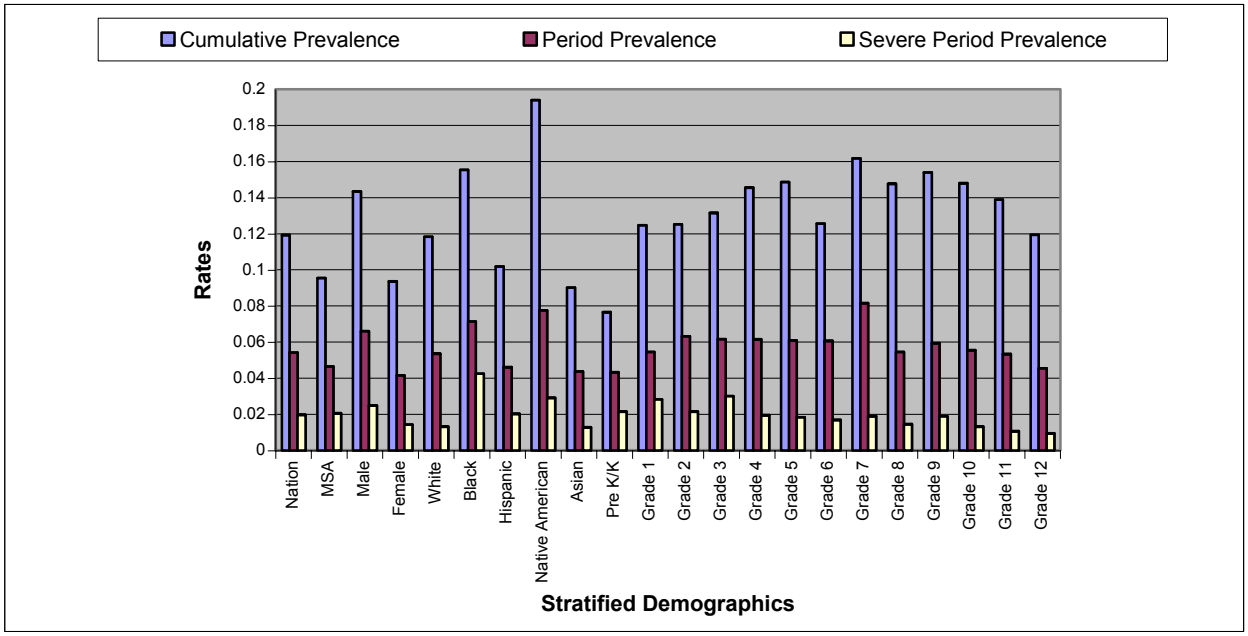



**Figure 2:** Cumulative Prevalence. Period Prevalence. and Severe Period Prevalence Rates for NHIS Modeled Strata.

declining. Rates of severe asthma tend to decline as grades increase.

The aforementioned GIS products are shown in Figures 3, 4, and 5. As with the neighborhood analysis shown in Figure 1, the estimates show higher rates on the south and west sides of Chicago, but the populations are made much more explicit in these products with some schools in these areas having estimates below the national level.

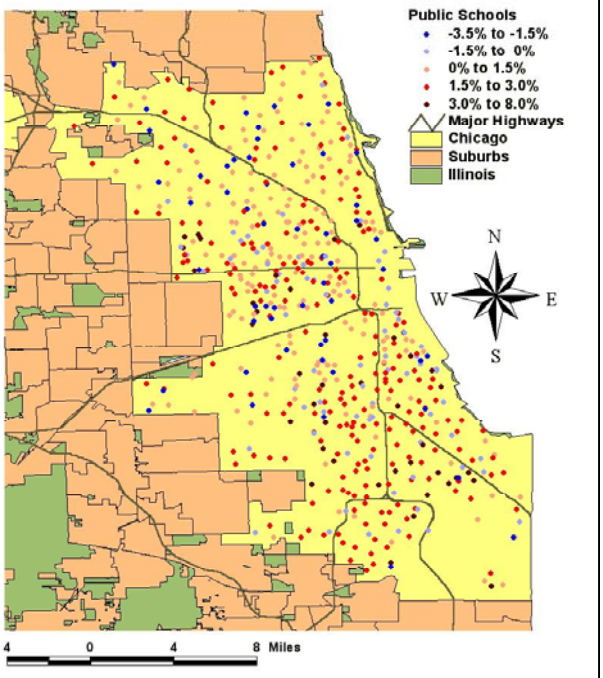



**Figure 3:** Difference Between Expected Cumulative Prevalence Rate of Asthma In Chicago Schools and the National Rate in 1998.

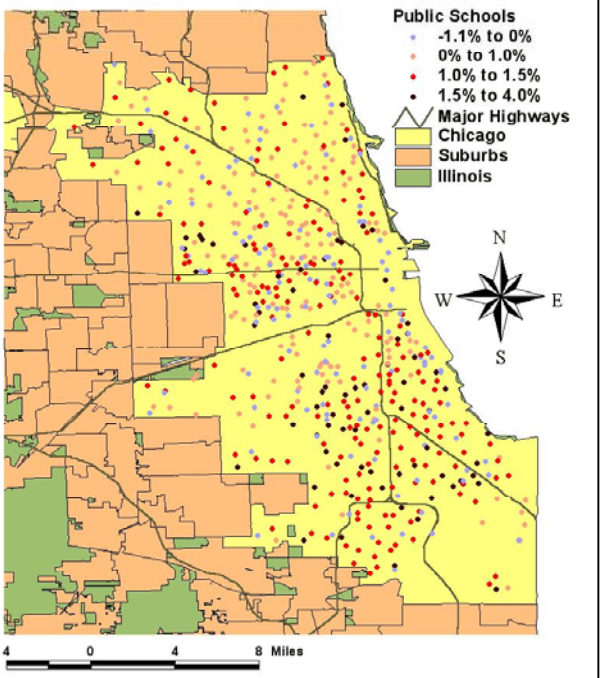



**Figure 4:** Difference Between Expected Period Prevalence Rate of Asthma In Chicago Schools and the National Rate in 1998.

## Next Steps

Four additional areas of exploration would be of interest from this work. First, more analysis should be conducted to establish the causes for the lower estimated cumulative and one year period prevalence rates of asthma in large Midwestern cities than that of the nation while the nation has a lower severe period prevalence. Only additional testing can identify the cause of these outcomes.

Second, the numbers generated in this analysis should be compared with school nurse data or other surveillance information on the number of students with asthma. Although asthma surveillance experts have seen school nurse data as an undercount of the true asthmatic population, it would be beneficial to compare empirical data with the modeled estimates for validation and possible intervention.

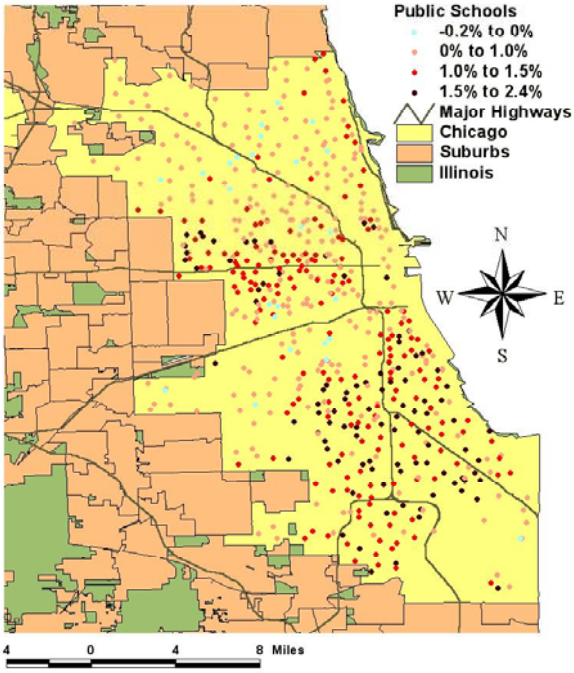



**Figure 5**: Difference Between Expected Severe Period Prevalence Rate of Asthma In Chicago Schools and the National Rate in 1998.

Third, the methods described in this paper could be used in a nationwide asthma prevalence estimation project. Of course the additional analysis would require a sufficient geocoded data set of the entire CCD and model refinements. US EPA is currently developing a national geocoded set of public schools, and it is hoped that this work will lead to interest in making the necessary refinements in the model to broaden the scale of the project.

Finally, and most importantly, the numbers reported in this paper are estimates derived from a modeling effort. Only an effective surveillance program will establish the true rates and variances in these small areas. Recent legislation has now made it possible for such a network to be developed nationally (CDC 2003). It will take time, but hopefully a much better understanding of the national asthma epidemic in children will come from this information network in the near future.

**References**

CDC's National Center for Environmental Health. CDC's Environmental Public Health Tracking Program - Background. Accessed 2/2003 at http://www.cdc.gov/nceh/tracking/background.htm (2003).

Geist M. Chicago Asthma Consortium. In: Report of the Second Annual Data Workshop. Asthma Prevalence accessed 2/2003 at http://chicagoasthma.org/2ndwrkshp.pdf (1999).

Murdock SH and Ellis DR. Applied Demography. Boulder: Westview Press, 1991.

Siegel JS. Applied Demography: Applications to Business, Government, Law and Public Policy. San Diego: Academic Press, 2002, 489-508.

Smith SK. Small-Area Demography. In: Encyclopedia of Population, Demeny P and McNicoll G, ed. Farmington Hills, MI: Macmillan Reference, 2003.

Wisconsin Department of Health & Family Services. Small Area Analysis in Southeastern Wisconsin, 1992-1994: Section 2. What is Small Area Analysis? Accessed 2/2003 at http://www.dhfs.state.wi.us/healthcareinfo/excerpts/saawg.htm (2003).

# Depiction of Population Characteristics in Areas Impacted by Industrial Source Emissions

Lawrence Lehrman, Arthur N. Lubin

*The purpose of this effort was to develop a set maps comparing the demographic characteristics of areas impacted by industrial source emissions. The demographic characteristics, derived from the 2000 Census, included age composition, racial composition, income distribution, sex composition and housing attributes. Industrial source emissions were from the most recent Toxics Release Inventory (Year 2000 TRI) . The resulting maps, portraying the demographic characteristics of areas nearest to reported emissions sources, can assist in environmental targeting, remediation and management. In addition to the maps and tables to demographically characterize the distribution of reported emissions, the results of statistical analyses are presented to demonstrate the significance and direction of relationships between demographics and reported emissions.*

## Analysis of Differential Levels of Environmental Burden

The purpose of this effort was to develop a set of analytical procedures for determining which block groups in census places (cities) and states have greater environmental burden and what are the certain demographic characteristics of areas with differential levels of environmental burden. The demographic variables selected were: proportions low income (ratio of income to poverty level); median family income, median house value, numbers of persons and population density, proportions racial/ethnic minorities, proportions less than one year of age, proportions aged 65 and above, proportions non-adults, and proportions of residences owned/rented. The data on demographics were obtained from the 2000 United States Census. A geographic block group (BG) is a cluster of blocks having the same first digit of their three digit identifying numbers within a census tract. BG's usually have from 250 to 550 housing units.

TRI (Toxics Release Inventory) air release data also were used for this analysis. The TRI data base involves stack and fugitive air emissions. Block Group TRI based estimated burdens were calculated using OPPT software which multiplies air stack and fugitive releases by EPA risk-based coefficients or weights. The burdens were calculated at the block group level by summing within a radius of the centroid of the block group the pounds released multiplied by the risk coefficients and dividing by the area. The result is toxicity released by unit of area. This provides a measure of the relative impact per individual.

In addition to providing maps and tables to show levels of comparative risk or burden, statistical analysis is used to assess the relationships between the level of environmental burden and the demographic characteristics and profile relative ranges of risk.

Analytical Procedures:

In order to clarify the procedure done, the steps involved in the process of calculating relative excess risk or burden from TRI data are summarized below:

1) OPPT software was used to multiply estimated stack and fugitive air emissions times EPA accepted risk-based coefficients or weights. The risk estimates were aggregated to the facility level.
2) Impacts of releases upon block groups were calculated by summing weighted TRI releases within a radius from the block group's center to the TRI facility.
3) We calculated the per capita average standard units of risk at the state and place levels by dividing the results of the prior to steps by the appropriate numbers of persons. This step is discussed in greater detail in the second paragraph of the purpose section of this writeup.
4. As previously mentioned, the results per block group were aggregated to calculate average standard units of risk at the place, state and block group levels.
5). Correlation along with statistical significance tests were used to demonstrate whether or not there are relationships between the environmental burdens or risk levels and the selected demographic characteristics as well as profile relative ranges of risk.
6) Tables were tabulated to show the statistical results from Step five.
7) Maps were produced for Region 5 showing the areas with their standard units of risk along with their demographics.

The steps are discussed in greater detail below:

Estimation of Ground Level Concentrations:
The results from EPA's Risk-Screening Environmental Indicator Model developed by the Office of Pollution Prevention and Toxics were used to estimate the relative continuous exposure from the TRI air releases. The model which has been assessed previously was used to provide a general determination of the relationships between source and reception distance. The reference for the OPPT model is:

Office of Pollution Prevention and Toxics. "EPA User's Manual for EPA's Risk-Screening Environmental Indicators Model: Version 1.02. (1988-1997 TRI Data, Air-Only Model). U.S. Environmental Protection Agency, Washington, D.C., 1999.

The initial step of the statistical analysis was to explore the characteristics of the data. This was followed by the use of correlational analysis to indicate whether or not there are substantial relationships between the demographics and risk as well as among the demographic characteristics.

Results

The analysis has been completed for 2000. The figures show the mapped 2000 burden results. In addition, the results of the associational analyses are shown. The results allow the identification of areas with high excess risk according to the TRI-based risk estimates and an assessment of whether the areas of high risk tend to have certain demographic attributes.

**Figure 1. Population Density Region 5**

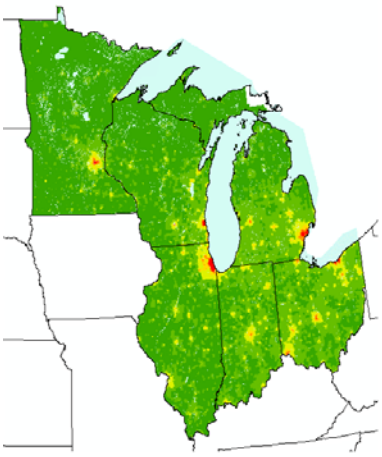
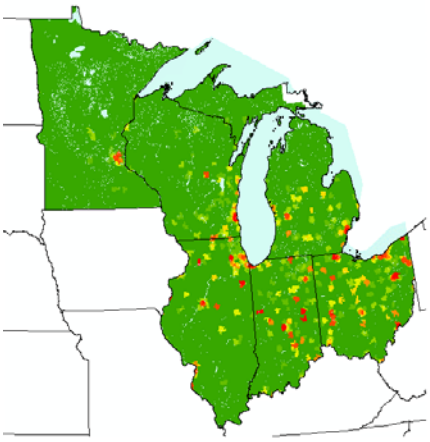
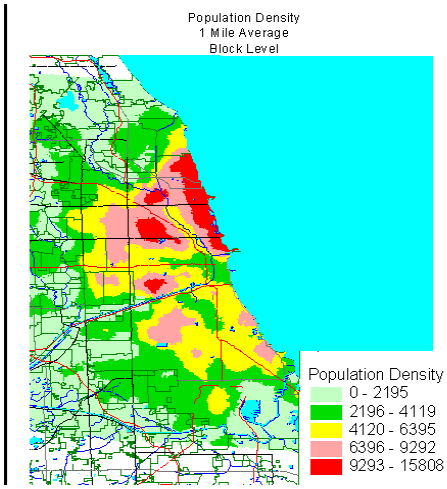


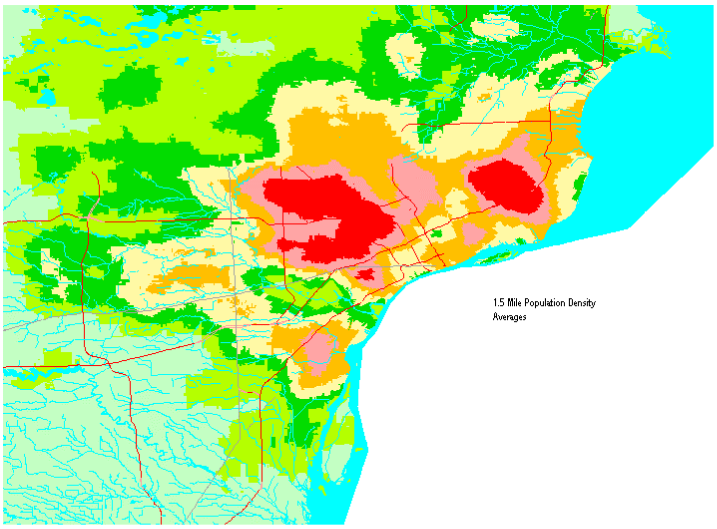
**Figure 2. Weighted TRI Air Releases Region 5**



## 3. Population Density Chicago




## Figure 4. Population Density Detroit



# Establishing Hazardous Emissions Limitations, Taming A Modern Hydra Using Statistical Methodology

Nelson Andrews, U.S. EPA

*As an environmental statistician within the EPA, my first environmental project was that of using statistical methodologies to develop point source MACT emissions limitations for hazardous air pollutants under guidance set forth by Congress and the Courts. This project has taught me that each environmental problem (or project) is likened to the hydra, a Greek mythological nine headed monster. When analyzing such environmental data, I discovered that one must be aware of many issues; these issues may raise their "heads" as legal impacts, confidentiality concerns, social impacts, political issues, legislative concerns, business concerns, enforcement issues, data integrity, and analytical correctness. First, I will discuss the specific programmatic concerns of implementing a statistical methodology to derive regulatory limits using MACT emissions data. In the first discussion, I discuss the design of the methodology and the statistical tools that are used in this project. I also touch upon the nature of the data and data challenges that are overcome through proper methodology design. Then I will conclude with a discussion of the issues that arise in establishing a statistical methodology to address the data integrity and analytical correctness issues while being cognizant of the other 'hydra-like' issues. For a statistical methodology to be cognizant of all these issues it must be reportable, repeatable, 'review able', reliable and modifiable. To tame such a modern 'hydra', the statistical team proposed a statistical methodology that begins with data identification and ends with management review. The implementation of the statistical analysis is done using "pseudo object oriented" design/programing (Visual Basic) and the presentation efficiency of electronic spreadsheets. Using statistical/programming modular design, this technical presentation illustrates the management of this specific environmental project (Setting Hazardous Emissions Limitations) by correlating individual designed modules to small issues that join together to encompass the more complex problems of this project.*

## Scope

Upon leaving the academic world and going to the world of private industry, I learned that nothing in the arena of data analysis is as clear and succinct as it was in school. But the one constant in the data analysis arena in private sector is that both, the customer and the supplier, are interested in the results of the analysis as it related to their costs. Hence, in private industry, money was the prevailing factor in determining the customer's or supplier's reaction to the results of the analysis. Making the transition from industry to the Environmental Protection Agency, I discovered that analytical results must address more than business impact issues; they must also address social issues, political issues, legal issues, economic issues and technical issues.

The project of establishing hazardous emissions standards began with the task of collecting the appropriate data. The data collection process is beyond the scope of this paper. However, the data collected helps define the size of the effort and to some extent it helps determine the

statistical methodology for the project. To establish standards, the engineers and participating institutions, collected hazardous waste combustion (HWC) data for Chlorine(HCL), Mercury (HG), Dioxin/Furans (DF), Particulate Matter (PM), Low Volatile Metals (LVM) and Semi Volatile Metals (SVM). In addition to these hazardous pollutants, six categories of combustion technologies. These combustion technologies are the following:

1. Incinerators
2. Light Weight Aggregate Kilns (LWAK)
3. Cement Kilns
4. HCL Production Furnaces
5. Coal Fired Boilers
6. Liquid Fuel Boilers.

Hence a minimum of 36 standards are to be developed using all combination of pollutants and combustion technologies described above.

## Issues And Concerns - Heads of The Hydra

Unlike the mythological hydra of Greek tragedies, the heads of the modern hydra referenced here are issues and concerns that must be addressed while performing statistical analysis of the data. Each head of the hydra represents an issue or concern, when considered in the analysis, may cause the analysis not to be straightforward and succinct as we desire. Each of these issues or concerns are briefly discussed below.

### (Social Issues)

The social issue is probably the easiest to understand. This issue addresses the negative and positive social impacts on the environment and people depending on what values are assigned to the standards. If the standards are not met or if the standards are set at a level that do not guard against emission levels that are harmful, then a negative social impact is realized. Such negative impacts are in the form of degraded health conditions, harm to the eco-system and a threat to future generation lifestyles. When the standards are set at reasonably low values and are being met by the various institutions, then society benefits in terms of increased health and a friendly environment that will produce in abundance.

### (Economic Impacts)

The economic impact of the standards are twofold; personal impact and business impact. The personal impact of the economy is the accounting side of the social issue associated with cost to individuals because of degenerating health when the standards are set too high or the institutions are not meeting the standards. When the personal cost is high, the economy is affected because such areas as productivity, resources and personal investment are negatively impacted. The business impact occurs when the standards are set too low and the investment by the business entities to meet these objectives are excessive. Such excessive spending in one area may translate to slower future growth for the business entity and hence a loss of current and future employment. To address these issues in the statistical analysis, it is imperative that the results of the statistical analysis takes a form that is useable by the economists to assess the economic impact of the resulting standards.

**(Legal Issues)**
With the legal decision of the D.C. Circuit Court of Appeals, dated July 24 2001, the matter of establishing the MACT standards has become a legal responsibility as well as a Congressional mandated responsibility as established in the Clean Air Act. The July 24th 2001 decision provided for a timeline for deriving the standards and cited specific sections within the Clean Air Act that provides specific methods for establishing a "floor". Section 112 (d) (3) of the Clean Air Act states the following definition for the "floor":

> "average emission limitation achieved by the best performing 12 percent of the existing sources (for which the Administration has emissions information) ....."

To adhere to the legal ruling set forth by the D.C. Circuit Court of Appeals, the statical analysis approach establishes limits (95%, 99%) on the average emission limitations achieved by the best performing 12 percent of the existing sources.

**(Political Issues)**
Each institution that generates hazardous waste emissions believes that their approach to establishing a standard is appropriate for their technology and industry. Unfortunately, there seem to be as many approaches as there are pollutant and combustion technology combinations. The combustion engineering team has screened the many approaches and have reduced the number to a manageable level. The statistical analysis will consider each of the approaches and generate the appropriate limits.

**(Confidentiality Concerns)**
As number are being generated, it has become increasingly clear that reviewers are anxiously awaiting the results. Since the final approach/methodology has not been decided, a code of confidentiality must be adhered to by all until the numbers can be released through the proper channels. Preliminary numbers that may change may cause some false conclusions and may cause some adverse political fall-out.

**(Enforcement Issues)**
The generation of standards that are not related to a simple and easily reproducible measurement of the emissions level is a disadvantage to the enforcement of the standard. Hence, the statistical analysis approaches have been consistent in defining the standard in an easily reproducible measurement.

**(Technical Issues)**
An important issue that resulted from the July 24th 2001 decision by the D. C. Circuit Court of Appeals is that variability issues must be considered. To address variability concerns, various subcategories of the data are analyzed and the issue of variances among these subcategories are assessed. To guard against nonsensical groupings of technology concerns, the combustion engineers provided the appropriate guidance and feedback concerning appropriate comparisons in the analysis.

**(Data Integrity)**
The issue of data integrity is a universal concern as it pertains to data analysis; regardless of the industry. To handle this issue a quality assurance plan and checklist is put into place to reduce the possibility of using data inappropriately throughout the analysis.

**(Analytical Correctness)**
Using the right tool for the right job is always a challenge, but it certainly simplifies the job when the right tool is used. In statistical analysis this is paramount to not making assumptions without using the appropriate statistical tests to verify the assumption. To guard against making bad assumptions, the statistical methodology for setting the MACT standards has statistical tests embedded throughout the analysis.

## Statistical Methodology

Before introducing the basic statistical methodology employed for this project, I will discuss the overall quality assurance/quality control approach used in the project analysis. The steps employed by project to provide assurance for such items as data integrity, technical correctness and analytical correctness are as follows:

1. Define and Describe Analytical Approach
2. Design Analytical Approach
3. Implement Analytical Approach
4. Review Data for Analytical Approach
5. Produce Limits Using Analytical Approach
6. Review Results of Analytical Approach
7. Repeat Above Steps (1....6) as Needed.

Step one(1) of the overall quality assurance/quality control approach is usually a present and discuss meeting where ideas are exchanged and a consensus is reached. In many instances, the combustion engineers will explain the technical issues and I will present and explain any new statistical concepts that will be uses. In many cases step two(2), will be modifications to a current methodology such as ranking by an upper limit or some other criterion.

The implementation of the analytical approach (step 3) involves the building of Visual Basic modules that will implement the analytical approach. Using modular designs (pseudo object oriented design), allows independent tasks to be separate and allows a single program to expand and still be readable and review-able. For step four (4), a summary sheet is produced for the engineers to review their area of expertise to make sure that data used in the analysis is the appropriate data. In steps five (5) and six (6), I perform the analysis and present the results to the engineers for review. If all goes well in the review, we have completed the analysis for one analytical approach. Otherwise, we continue steps 1 through 6 until we get it right.

A flow diagram showing the initial methodology for the analysis is shown in figure 1.

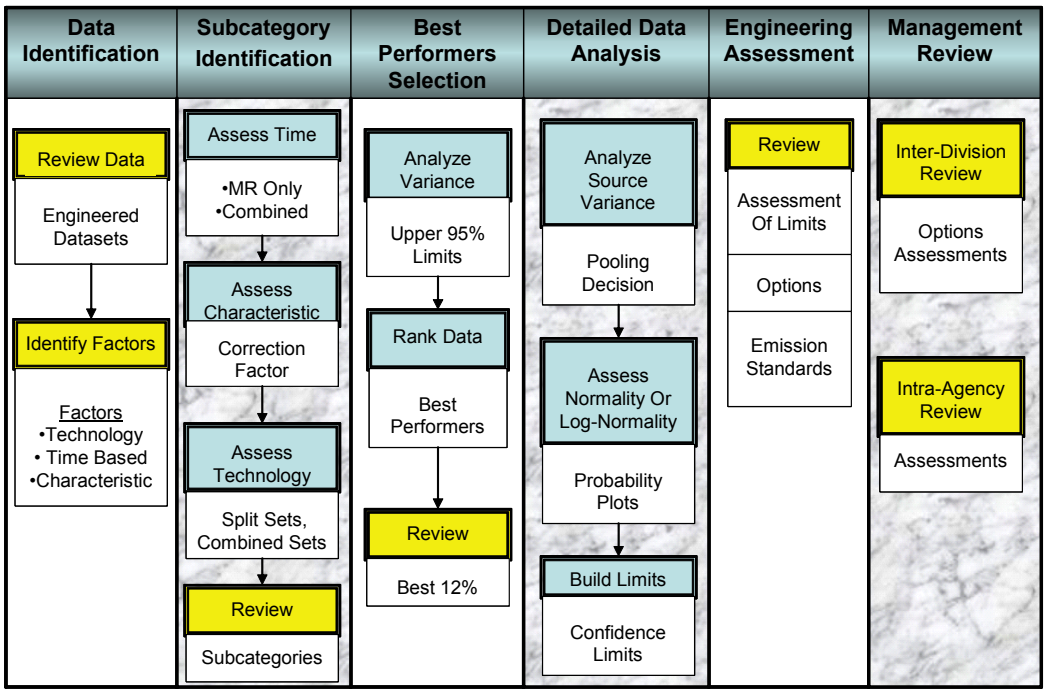



**Figure 1. Statistical Methodology For Deriving MACT Limits**

The methodology, illustrated in Figure 1, is an evolving methodology and changes as new discussions with the engineers occur. The statistical methodology begins with data identification, then subcategory identification. The subcategory identification looks at factors such as time, ratings (Worst Case, Normal or In Between) and technology. It examines these factors to determine if we should separate the data using the different levels of these factors. In the case of time, if there is a significant downward trend in the emission's level over time, then we only include the most recent data in the analysis.

The next two phases of the analysis "Best Performers Selection" and "Detailed Data Analysis" represent the core of the statistical analysis. Remembering that the legal decision handed down by the D. C. District Courts, we must determine the best performers (12% or a minimum of 5 when possible). Upon determining the best performers, we then proceed to compute the stochastic limits above the average of the best performers of a three run average for a future stochastic event.

Once the limits are computed, we then enter the review phases "Engineering Assessment" and "Management Review".

## Results

The results of the analysis are summarized in a single spreadsheet for a given approach. A typical spreadsheet for a given approach is shown in Figure 2.

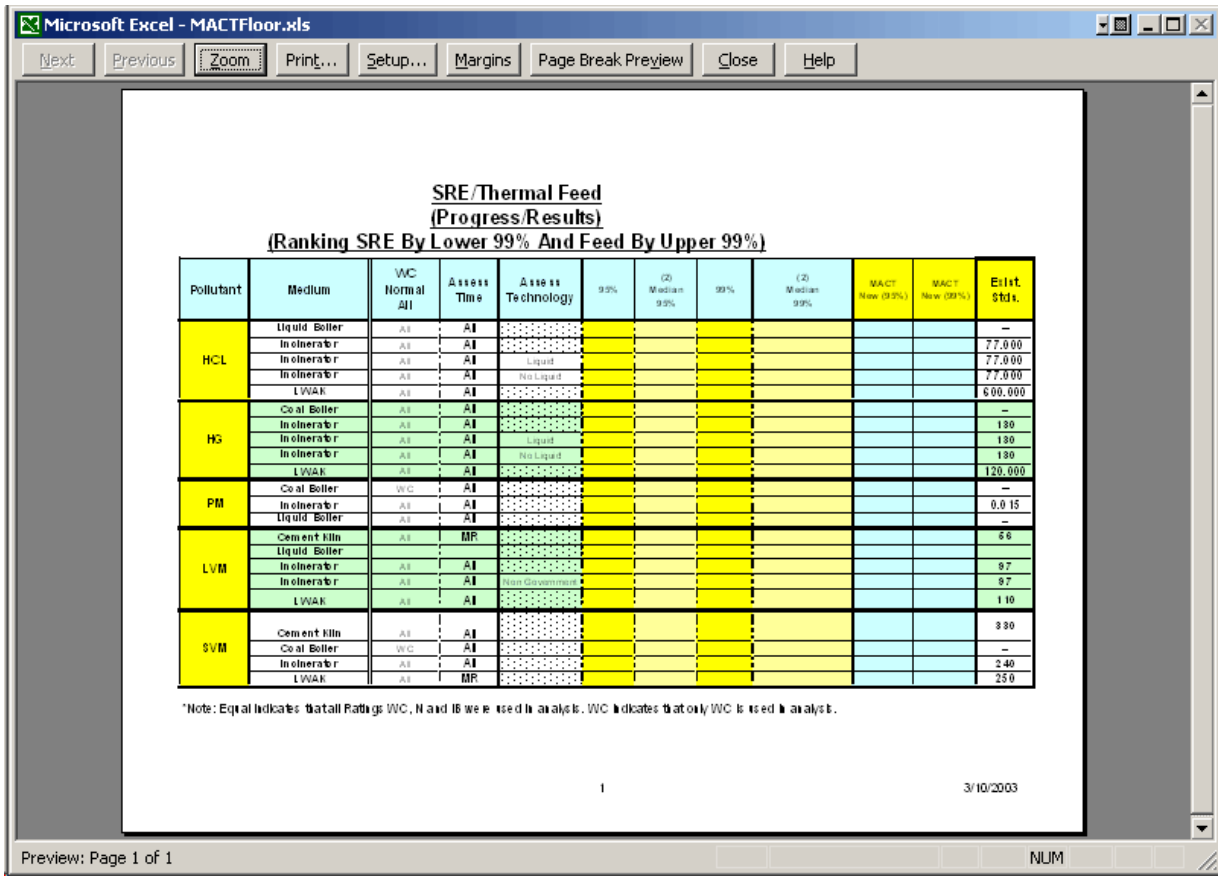



**Figure 2. Example of Summary Sheet For A Single Approach For Deriving MACT Limits**

# Estimation of the Exposure Point Concentration Term Using a Gamma Distribution

Anita Singh
Lockheed Martin Environmental Services
1050 E. Flamingo Road, Suite E120, Las Vegas, NV 89119
Phone: 702-897-3234, Email: asingh@lmepo.com

John M. Nocerino
Unites States Environmental Protection Agency
National Exposure Research Laboratory
Characterization and Monitoring Branch
P.O. Box 93478, Las Vegas, NV 89193-3478

Ashok K. Singh
Department of Mathematical Sciences
University of Nevada, Las Vegas, NV 89154

*In Superfund and RCRA projects of the United States Environmental Protection Agency (U.S. EPA), cleanup, exposure, and risk assessment decisions are often made based upon the mean concentrations of the contaminants of potential concern (COPC). A 95% upper confidence limit (UCL) of the population mean, $\mu$, is used to: estimate the exposure point concentration (EPC) term, determine the attainment of cleanup standards, estimate background level contaminant concentrations, or compare the soil concentrations with the site-specific soil screening levels. It is, therefore, important to compute an accurate and stable 95% UCL of $\mu$ from the available data. The formula for computing a UCL depends upon the data distribution. Typically, environmental data are positively skewed and can quite often be modeled by lognormal or gamma distributions. Due to computational ease, the lognormal distribution is usually used to model positively skewed data sets. However, the use of a lognormal model for an environmental data set unjustifiably inflates the minimum variance unbiased estimate of $\mu$ and its UCL to levels that may not be applicable in practice. In this paper, we propose the use of a gamma distribution to model positively skewed data sets. The objective of the present work is to study procedures which can be used to compute a stable and accurate UCL $\mu$ based upon a gamma distribution. Several parametric and non-parametric (e.g., the standard bootstrap, the bootstrap-t, Hall's bootstrap, and the Chebyshev inequality) methods of computing a UCL of an unknown $\mu$ have also been considered. Monte Carlo simulation experiments have been performed to compare the performances of those methods. The comparison of the various methods has been evaluated in terms of the coverage (confidence coefficient) probabilities achieved by those various UCLs. Based upon this study, recommendations are made about the computation of a UCL of an unknown $\mu$ for skewed data distributions originating from various environmental applications.*

**The Gamma Distribution and Computation of UCL of the Population Mean, μ**

A random variable, X (*e.g.*, lead concentrations), follows a gamma distribution, *G(k,θ),* with parameters $k > 0$ and $\theta > 0$, if its probability density function is given by the following equation:

$$f(x;k,\theta) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta}; \quad x > 0, \tag{1}$$

and zero otherwise. The mean, variance, and skewness of a gamma distribution are: mean = $\mu = k\theta$, variance = $\sigma^2 = k\theta^2$, and skewness = $2/\sqrt{k}$. Note that, as *k* increases, the skewness decreases, and, consequently, the gamma distribution starts approaching a normal distribution for larger values of *k* (*e.g.*, $k \geq 6$ - 8). It is observed that for larger values of *k*, the UCL based upon a gamma distribution and a UCL based upon a normal distribution are in close agreement (Singh *et al.,* 2002). Several other UCL computation methods have also been studied via Monte Carlo experiments (Singh *et al.*, 2002). Those methods include: the Student's t-statistic, the modified t-statistic, the adjusted central limit theorem (CLT), Land's H statistic based UCL, the Chebyshev inequality-based UCL, and UCLs based upon bootstrap procedures (Efron, 1982; Hall, 1992). In this brief article, the computation of those methods have been illustrated via some examples. For details, the reader is referred to Singh *et al*., (2002). Gamma UCLs are computed using the program, Chi_test, and the software, ProUCL (EPA 2002a, EPA 2002b), has been used to compute all of the other UCLs.

The first step is to test if a given data set follows a gamma distribution. Several goodness-of-fit tests for the gamma distribution are available in the statistical literature (D'Agostino and Stephens, 1986). Those tests are based upon a chi-square goodness-of-fit test, the Kolmogorov-Smirnov D test statistic, and the Anderson-Darling $A^2$ test statistic. If the data follow a gamma distribution, then the procedure described as follows can be used to compute a UCL of μ. The gamma UCL of μ depends upon the maximum likelihood estimate (MLE) of *k,* which is quite complex and requires the computation of digamma and trigamma functions (Singh *et al*., 2002). Given a random sample, $x_1, x_2, \ldots, x_n$, of size n from a gamma, *G(k,θ)*, distribution, it can be shown that $2n\bar{X}/\theta$ follows a Chi-square distribution, $\chi^2_{2nk}$, with *2nk* degrees of freedom (ν). Using this relationship, an approximate *(1-α)* 100% UCL of μ can be computed as follows:

$$\textit{Approximate-}UCL = 2n\hat{k}^*\bar{x} / \chi^2_{2n\hat{k}^*}(\alpha) \tag{2}$$

Where $\chi^2_{\upsilon}(\alpha)$ denotes the α cumulative percentage point (lower) of the chi-square distribution, and $\hat{k}^*$ is the bias-adjusted estimate of *k*. Simulation studies (Singh *et al.*, 2002) suggest that an approximate gamma 95% UCL thus obtained provides the specified coverage (95%) of μ as *k* approaches 0.5. Thus when $k \geq 0.5$, one can use the approximate 95% UCL of the mean to estimate the EPC term. This approximation is good even for smaller (*e.g.*, n = 5) sample sizes, as shown by Singh *et al.* (2002). For smaller values of *k* (*e.g.*, $k < 0.5$), one can use the following adjusted *(1-α)* 100% UCL (*e.g.*, 95% UCL with $\alpha = 0.05$) of the gamma mean, $\mu = k\theta$.

$$Adjusted\text{-}UCL = 2n\hat{k}^{*}\bar{x} \,/\, \chi^{2}_{2n\hat{k}^{*}}(\beta), \qquad (3)$$

where β is the adjusted level of significance associated with the specified level of significance, α. It is observed (Singh *et al.*, 2002) that, except for highly skewed ($k < 0.15$) data and samples of a small size (*e.g.*, n < 10), the adjusted gamma 95% UCL given by equation (3) provides the specified 95% coverage of μ. It is also noted that for highly skewed ($k < 0.15$) data sets of small sizes, except for the 95% H-UCL, the coverage probability provided by the adjusted gamma 95% UCL is the highest (among all of the other methods) and is close to the specified level, 0.95. However, for those highly skewed data sets of small sizes, the H-UCL results in unacceptably large values of the 95% UCL, as shown in examples 1 - 3. For values of $k > 0.15$, the coverage of 0.95 is always approximately achieved by the adjusted 95% gamma UCL, as shown by Singh *et al.* (2002).

**Example 1**. A data set of size n =15 was generated from a gamma, $G(0.2, 100)$, distribution with the true population mean, μ = 20, and the skewness = 4.472. The data are: 0.7269, 0.00025, 0.0000002548, 0.9510, 0.000457, 32.5884, 0.02950, 1.6843, 3.3981, 170.4109, 59.8188, 0.00042, 0.8227, 0.00726, and 2.1037. The data set consists of very small values (such as non-detects) as well as some large values. Those types of data sets often occur in environmental applications. Using the Shapiro-Wilk's W-test (EPA, 2002b), it is concluded that the data follow a lognormal model. The standard deviation (sd) of the log-transformed data is large, sd = 5.618; therefore, the H-statistic based UCL of the mean becomes impractically large ($5.4 \times 10^{13}$). The bias-corrected MLEs (Singh *el al.*, 2002) of *k* and θ are 0.165 and 109.939 with adjusted $\hat{v}^{*} = 4.958$. The various UCLs are summarized in the following table.

| **UCL Computation Method** | **95% UCL of Mean** |
|---|---|
| approximate gamma UCL | 79.968 |
| adjusted gamma UCL | 98.139 |
| UCL based upon the Student's t-statistic | 38.778 |
| UCL based upon the modified t-statistic | 40.356 |
| UCL based upon the adjusted CLT | 47.537 |
| UCL based upon the H-statistic | 5.4E+13 |
| UCL based upon the Chebyshev inequaltiy | 69.171 |
| UCL based upon the standard bootstrap | 36.889 |
| UCL based upon the bootstrap-t | 102.392 |
| Hall's bootstrap UCL | 114.252 |

Since the H-UCL is larger than the maximum observed value (170.41), using the U.S. EPA RAGS guidance document (EPA 1992), one would use this maximum value as an estimate of the EPC term. Simulation results (Singh *et al.*, 2002) suggest that for n = 15 and an estimate of $k = 0.165$, the adjusted gamma 95% UCL provides the specified 95% coverage to μ. Therefore, the adjusted UCL (98.14) provides a more accurate average estimate of the EPC term.

**Example 2**. A data set of size n = 15 was generated from a gamma distribution with $k$ = 0.5 and $\theta$ =100 with $\mu = k\theta = 50$ and skewness = 2.828. The data are: 343.31, 102.44, 0.33, 1.42, 13.17, 439.59, 130.66, 158.0, 70.65, 25.05, 144.84, 63.65, 62.50, 11.58, and 1.097. Using the Shapiro-Wilk's test (EPA 2002b), it is concluded that the hypothesis that the data follow a lognormal distribution cannot be rejected. The sample mean = 104.553. The bias-corrected estimates of $k$ and $\theta$ are 0.462 and 226.473, respectively. For the chi-square distribution, the adjusted $\upsilon^* = 2n\hat{k}^* = 13.85$. For $\alpha$ = 0.05 and n = 15, the adjusted critical probability level, $\beta$ = 0.0324. The 95% UCLs of the mean obtained using the various methods described above are given below.

| **UCL Computation Method** | **95% UCL of Mean** |
|---|---|
| approximate gamma UCL | 223.879 |
| adjusted gamma UCL | 247.257 |
| UCL based upon the Student's t-statistic | 163.413 |
| UCL based upon the modified t-statistic | 165.92 |
| UCL based upon the adjusted CLT | 175.596 |
| UCL based upon the H-statistic | 5687.383 |
| UCL based upon the Chebyshev inequality | 250.22 |
| UCL based upon the standard bootstrap | 158.798 |
| UCL based upon the bootstrap-t | 223.665 |
| Hall's bootstrap UCL | 461.795 |

Again, note that the H-UCL (5687.38) is much higher than the UCLs obtained using any of the other methods. Simulation results suggest that for n=15 and an MLE of $k$ = 0.462, both approximate as well as the adjusted 95% UCLs based upon the gamma model provide the specified 95% coverage of $\mu$. Also note that the Chebyshev inequality 95% UCL is very close to the adjusted gamma 95% UCL. Any of those three methods may be used to compute the 95% UCL.

**Example 3.** A sample of size n = 15 was generated from the lognormal distribution with the parameters (of log-transformed variable) mean = 5 and sd = 2. The true meanof the lognormal distribution is 1096.6, the coefficient of variation (CV) is 7.32, and the skewness is 414.4. The generated data are: 47.42, 2761.51, 2904.26, 6928.33, 14.73, 7.67, 73.36, 2843.79, 151.71, 103.52, 14.8, 37.32, 24.74, 658.04, and 110.42. Using the Anderson-Darling and the chi-square goodness-of-fit tests, it is concluded that an approximate gamma distribution can also be used to model the distribution of this data set. The bias-adjusted estimates of $k$ and $\theta$ are 0.321 and 3466.301, respectively. The 95% UCLs computed from the various methods are given below. Notice that the H-UCL is more than 5 times higher than the maximum concentration in the sample, and more than 10 times higher than all of the other UCLs. Using the simulation results (Singh *et al.*, 2002), it is observed that for an estimate of $k$ = 0.321 and n = 15, the adjusted gamma 95% UCL = 3276.40 provides the specified 95% coverage to $\mu$. Therefore, instead of

using the maximum concentration, the adjusted gamma UCL may be used as an estimate of the EPC term. It should be pointed out that the H-statistic based UCL of μ results in unjustifiably inflated values even when the data are generated from a lognormal distribution.

| **UCL Computation Method** | **95% UCL of Mean** |
|---|---|
| adjusted gamma UCL | 3276.40 |
| UCL based upon the Student's t-statistic | 2005.97 |
| UCL based upon the modified t-statistic | 2053.46 |
| UCL based upon the adjusted CLT | 2251.31 |
| UCL based upon the H-statistic | 37726.46 |
| UCL based upon the Chebyshev inequality | 3324.25 |
| UCL based upon the standard bootstrap | 1917.43 |
| UCL based upon the bootstrap-t | 2541.43 |
| Hall's bootstrap UCL | 2305.17 |

Continuing with this example, suppose another sample (observation) is collected and it is below the detection limit (DL) of the instrument. Let DL = 10, and following standard practice, this value is replaced by DL/2 = 5. Typically, one would expect that this additional non-detect observation would result in a reduction in the average value and the associated UCL. The UCLs calculated using this sample of n = 16 observations are given as follows.

| **UCL Computation Method** | **95% UCL of Mean** |
|---|---|
| adjusted gamma UCL | 3013.75 |
| UCL based upon the Student's t-statistic | 1883.92 |
| UCL based upon the modified t-statistic | 1929.28 |
| UCL based upon the adjusted CLT | 2122.80 |
| UCL based upon the H-statistic | 40313.16 |
| UCL based upon the Chebyshev inequality | 3134.05 |
| UCL based upon the standard bootstrap | 1795.20 |
| UCL based upon the bootstrap-t | 2369.23 |
| Hall's bootstrap UCL | 2167.02 |

The UCLs computed using all but the H-statistic based UCL decreased with the addition of one non-detect value; the H-statistic based formula, however, resulted in a much larger UCL. This is unacceptable and impractical behavior of the H-statistic based UCL of μ.

**3.0 Recommendations**

Positively skewed data sets can be modeled by more than one statistical distribution. Due to the computational ease of working with a lognormal model, users often choose the lognormal distribution. For small data sets, it is not easy to distinguish between a gamma model and a lognormal distribution. As noted, there are some fundamental problems associated with the use

of a lognormal distribution. This is especially true when the skewness is high and the sample size is small. The use of a lognormal model unjustifiably inflates the mean and the associated UCL; therefore, its use in environmental applications should be avoided. Since the H-UCL becomes unrealistically large, the Max-test is sometimes used (EPA, 1992) as an estimate of the EPC term. It is shown (Singh *et al.*, 2002) that for highly skewed ($k < 0.25$) data sets of small sizes ($n < 10$), the Max-test does not provide the specified 95% coverage to the means of gamma populations, and for larger samples, the Max-test overestimates of the EPC term. Since the EPC term represents an average exposure concentration in an area, it should be estimated by a conservative average value, such as a 95% UCL of the mean. In this paper, we have introduced the gamma distribution to model highly skewed data sets originating from various environmental applications. The use of the gamma distribution results in practical and reliable UCLs of $\mu$. Simulation results discussed by Singh *et al.* (2002) suggest that the adjusted gamma 95% UCL approximately provides the specified 95% coverage of $\mu$ for data sets with the shape parameter, $k$, exceeding 0.15. Thus, it is recommended that the user perform a goodness-of-fit test to see if the data follow a gamma distribution. If the data follow a gamma distribution, then the user should compute a UCL of the mean based upon a gamma model. It is shown that the approximate and the adjusted gamma UCLs behave in a stable manner. For estimated values of the shape parameter with $k \geq 0.5$, one can use the approximate gamma UCL as an estimate of the EPC term, and for values of $k < 0.5$, one can use the adjusted gamma UCL of the mean. For data sets which cannot be modeled by an approximate gamma distribution, one can use a UCL based upon the Chebyshev inequality or the bootstrap t-procedure, provided no outliers are present. Outliers should be removed before computing the UCL of the mean based upon the bootstrap-t procedure. Those two procedures generally result in conservative, but reasonable, estimates of the EPC term provided, no outliers are present.

**References**


1. D'Agostino, R.B. and Stephens, M.A. (1986). Goodness-of-Fit-Techniques. Marcel Dekker, Inc. New York.
2. Efron, B. (1982), The Jackknife, the Bootstrap, and Other Resampling Plans, Philadelphia: SIAM.
3. EPA (1992), Supplemental Guidance to RAGS: Calculating the Concentration Term, Publication EPA 9285.7-081, May 1992.
4. EPA (2002a), Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites, OSWER 9285.6-10, December 2002.
5. EPA (2002b), ProUCL Version 2.1 Statistical Software, National Exposure Research Laboratory, EPA, Las Vegas Nevada, April 2002.
6. Hall, P. (1992), On the Removal of Skewness by Transformation. Journal of Royal Statistical Society, B 54, 221-228.
7. Singh, A., Singh, A. K., and Iaci, R. (2002), Estimation of the Exposure Point Concentration Term Using a Gamma Distribution. Technology Support Center Issue Paper, EPA/600/R-02/084. October 2002.

# Integration of Spatial Data: Methods Evaluation with Regard to Data Issues and Assessment Questions

Michael O'Connell[1], Elizabeth R. Smith[2], R.V. O'Neill[3], Liem T. Tran[4], and Nick Locantore[1]

[1]Waratah Corporation, Durham, NC; [2]US EPA, National Exposure Research Laboratory, Research Triangle Park, NC; [3]TN and Associates, Oak Ridge, TN; [4]Pennsylvania State University, College Park, PA

*EPA's Regional Vulnerability Assessment (ReVA) Program is developing and demonstrating approaches to assess current and future environmental vulnerabilities at a regional scale. An initial effort within this research program has been to develop and evaluate methods to synthesize existing spatial data on resource condition and sensitivity, and estimated stressor distributions to facilitate decision-making on alternative environmental policies or risk management strategies. A total of 9 methods, ranging from simple spatial overlays to estimates of changes in multivariate state space, have been tested with regard to sensitivity to data issues such as skewed distributions, continuous versus discontinuous data, and imbalance of indicators or metrics (e.g. a large amount of terrestrial data versus a small amount of aquatic data), as well as suitability for addressing different assessment questions. Utilizing available data for the Mid-Atlantic region over a total of 73 variables, testing was also done to identify whether different integration results were similar, or whether individual methods provide unique information and should be used in concert. The results of this analysis identifies potential limitations of methods due to data structure and suggests that assessment of vulnerabilities can best be accomplished using a suite of methods that rank on condition, vulnerability (risk of future damage), and risk management feasibility.*

*This work has been funded wholly or in part by the United States Environmental Protection Agency under contract number 68-C98-187 with TN and Associates, cooperative agreement number R-82880301 with Pennsylvania State University, and contract number ISE00029 (COMMITS) with Waratah Corporation. It has been subjected to Agency review and approved for publication.*